ENTATION PAGE

Form Approved
OMB No. 0704-0188

1 AGENCY USE ONLY (Leave blank)	2 REPORT DATE	3 REPORT TYPE AND DATES COVERED
	May 1991	presentation/paper

4 TITLE AND SUBTITLE

A SCENARIO-BASED METHODOLOGY FOR EVALUATING USER INTERFACE FUNCTIONALITY ON A DATABASE RETRIEVAL TASK

5 FUNDING NUMBERS

In-house

6 AUTHOR(S)

K. Fernandes

7 PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Ocean Systems Center
San Diego, CA 92152–5000

8 PERFORMING ORGANIZATION REPORT NUMBER

9 SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Naval Ocean Systems Center
San Diego, CA 92152–5000

10 SPONSORING/MONITORING AGENCY REPORT NUMBER

11 SUPPLEMENTARY NOTES

12a DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b DISTRIBUTION CODE

13 ABSTRACT (Maximum 200 words)

The current study developed a methodology that attempts to link ease and accuracy in user interface operation to system output and utility. A problem-solving task is broken into four components—problem, interface, output, and user decision. With this methodology, a problem is presented and then data are collected on the operation of each of the remaining components. These data can be used to determine (1) if a user interface produces the system output and/or user decision expected or desired and (2) if alternative interface designs have a differential impact on the system output and/or user decision.

Published in *Proceedings of the Department of Defense Human Factors Engineering Technical Group*, Nov 1990.

14 SUBJECT TERMS			15 NUMBER OF PAGES
			16 PRICE CODE
17 SECURITY CLASSIFICATION OF REPORT	18 SECURITY CLASSIFICATION OF THIS PAGE	19 SECURITY CLASSIFICATION OF ABSTRACT	20 LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAME AS PAPER

21a NAME OF RESPONSIBLE INDIVIDUAL	21b TELEPHONE *(include Area Code)*	21c OFFICE SYMBOL
K. Fernandes	(619) 553–9224	Code 442

A SCENARIO-BASED METHODOLOGY FOR EVALUATING USER INTERFACE FUNCTIONALITY ON A DATABASE RETRIEVAL TASK

Kathleen Fernandes, Ph.D.
Naval Ocean Systems Center

BACKGROUND

Although analytical and theoretical guidelines are available for designing effective user interfaces, little work has been done to actually test these guidelines against interface performance. As a result, evaluative (and comparative) data are lacking about how the usability of an interface contributes to overall system effectiveness. Additionally, the selection of a user interface is an important consideration in system design since the choice has been shown to systematically influence user decision making. A methodology is clearly needed, therefore, to select the most appropriate user interface for a decision support system, both as the system currently function and as it evolves in complexity.

A current example of a decision support system that addresses complex problem solving tasks is the Force Requirements Expert System (FRESH), one of several systems being developed by the Fleet Command Center Battle Management Program (FCCBMP). FRESH is designed to monitor the readiness of Fleet forces, identify force deficiencies, and recommend alternatives for responding to deficiencies and meeting new mission requirements. Access to FRESH's functions is provided through Natural Language Menu, a menu-based, near-natural language user interface that allows users to build queries by selecting words and phrases from an array of menus that appear on the FRESH screen. In addition, the potential applicability of natural language interfaces to FCCBMP systems is being examined through experimentation with two commercially licensable software products. Both interface products are typed-entry interfaces that allow the user to freely create a query in English, thus permitting queries to be phrased in the manner most meaningful to the individual user.

A METHODOLOGY FOR INTERFACE EVALUATION

The current study developed a methodology that attempts to link ease and accuracy in user interface operation to system output and utility. A problem-solving task is broken into four components --- problem, interface, output, and user decision. With this methodology, a problem is presented and then data are collected on the operation of each of the remaining components. These data can be used to determine (1) if a user interface produces the system output and/or user decision expected or desired and (2) if alternative interface designs have a differential impact on the system output and/or user decision.

Although the methodology has potential applicability for evaluating user interface effectiveness on a range of problem solving tasks, the current study used a database retrieval task

as the initial application for testing the methodology. Natural Language, one of the two natural language interfaces being considered for FCCBMP application, was chosen because of its availability at the time of the study. The goals of the test were modest: (1) create a set of domain-specific scenarios that would provide a fair and reasonable test of interface usability, (2) determine if the task of generating database queries in response to these scenarios produces sufficient individual variability to warrant analysis, and (3) develop and apply a set of performance measures to use in assessing user and system operations within the interface component of the methodology.

APPROACH

Although the current study focused on a single user interface, scenarios were developed that were appropriate to all three FCCBMP interfaces so that the scenarios could be used if desired to compare performance across interfaces. The FRESH database was reviewed to identify the set of database tables that were common to all three interfaces. Available transcripts of database queries were reviewed, and a list of subject areas and generic query forms (i.e., phrased without reference to a specific ship, ship attribute, or subject area) was produced that could be used to generate scenarios appropriate across all three interfaces. The query forms were then reviewed with researchers who were customizing Natural Language for FCCBMP application to ensure that the forms provided a fair and meaningful test of the interface. Finally, specific scenarios for the current study were developed for each of the 26 query forms.

Seven civilian employees at the Naval Ocean Systems Center participated in the study during March and April 1990. After receiving an explanation of how to use the interface, the participants were given the set of scenarios and asked to build a query that would provide the information called for in the scenario. The participants could enter as many queries as desired until they felt they had obtained the information requested.

Session logs generated by the interface recorded the queries, paraphrases, and database commands (in Structured Query Language [SQL] form) for each participant. Following a review of these logs, a set of codes was developed to measure the extent to which (1) a query requested the information called for in the scenario and (2) the interface was able to (a) understand the query (i.e., generate a meaningful paraphrase) and (b) construct an accurate SQL that would retrieve the appropriate information. The queries, paraphrases, and SQLs entered by the participants were coded independently by two raters and reviewed with the researchers familiar with the interface to ensure coding accuracy.

The query-building skills of participants were evaluated by calculating the percentage of "good" queries (i.e., linguistically accurate and at least partially responsive to the requirements of the scenario) that were produced for each scenario. The primary measures of the query-understanding skills of the interface were obtained by calculating (1) the percentage of accurate

paraphrases in the set of "good" queries and (2) the percentage of accurate SQLs in the set of "good" queries with accurate paraphrases.

RESULTS AND DISCUSSION

The set of domain-specific scenarios that were developed were effective in eliciting a range of database queries from the participants in the study. Eighty percent of the queries generated were judged as "good," i.e., reasonable candidates for the interface to understand and execute. In addition, participants differed in their ability to build "good" queries, suggesting that there was sufficient individual variability to warrant analysis of task performance. The measures used to assess interface functionality indicated Natural Language had limited ability to understand and execute the queries built by the participants. The interface produced an accurate paraphrase for only about half of the "good" queries and an accurate SQL command for only about half of these queries (i.e., about 25 percent of the "good" queries).

A range of errors and inconsistencies in interface operation was identified in the data generated by the participants. For example, when producing a paraphrase, the interface sometimes translated incorrectly a key term in a query, was unable to recognize synonyms or qualifiers, and could not understand the range of linguistic constructions posed by the participants in their queries. The problems in SQL generation were due primarily to the inability of the interface to generate SQL commands for "good" queries that had been paraphrased accurately.

Participants' reactions to Natural Language were generally negative. They were dissatisfied with the overall performance of the interface and frustrated by (1) its inability to understand what they considered to be an acceptable query and (2) the lack of informative feedback from the interface when it was unable to produce a paraphrase or SQL command. Participants' comments during testing suggested they were both performing the task of generating queries in response to the scenarios and trying out different query-building strategies to determine how to work around the idiosyncracies being exhibited by the interface.

The queries generated in the current study were provided to the researchers who were adapting Natural Language for FCCBMP application so that the errors and inconsistencies identified could be corrected in later versions of the interface software. After obtaining a new release of the software and updating the interface's knowledge base, the researchers reran the query set from the current study. The performance measures for the revised software were 72 percent accurate paraphrases and 65 percent accurate output for the set of "good" queries that had been generated by participants. The researchers were able to use the performance data from the initial and revised versions of the software to demonstrate the extent to which interface functionality could be improved. In addition, the researchers used the data to determine the marginal gains in performance that might be achieved with further revisions to the software and to

decide if these improvements should be made given the costs associated with software revision.

CONCLUSIONS

Although exploratory and with limited goals, the current study demonstrated a scenario-driven methodology can provide meaningful and useful data for evaluating interface effectiveness in a database retrieval task. Quantitative operator performance data were generated from a realistic, operationally-relevant task. In addition, meaningful measures of performance improvement were generated that can be used to evaluated progress made in the software development process. Although participants were able to translate the information requirements of the scenarios into a series of database queries, the interface was only modestly successful in understanding the queries and constructing database commands to retrieve the information desired. Additional research should use the methodology to compare alternative interface designs, examine the operation of other components of the methodology, and determine its utility in evaluating interface effectiveness on a range of problem solving tasks.